Robust Image Forgery Detection and Localization Framework using Vision Transformers (ViTs)

Mahesh Enumula, M. Giri, V. K. Sharma

***Abstract** Image forgery detection has become increasingly critical with the proliferation of image editing tools capable of generating realistic forgeries. Traditional deep learning approaches, such as convolutional neural networks (CNNs), often struggle with capturing global dependencies and subtle inconsistencies across larger image contexts. To address these challenges, this paper proposes a novel Vision Transformer(ViT)-based framework for robust image forgery detection and localization. Leveraging the self-attention mechanism of transformers, our approach effectively models long-range dependencies and detects even subtle tampered regions with high precision. The proposed framework processes images as patch embeddings, extracting both local and global features, and outputs a detailed forgery map for accurate localization. We evaluate our method on multiple benchmark datasets containing diverse forgery types, including splicing, cloning, and inpainting. Experimental results demonstrate that the ViT-based model outperforms state-of-the-art CNN and GAN-based methods, achieving superior accuracy, precision, and recall. Additionally, qualitative analyses highlight its capability to localize forgeries in complex scenarios. The results underscore the potential of Vision Transformers as a powerful tool for advancing the fieldof image forgery detection.*

***Index Terms:** Image Forgery Detection, Vision Transformers (ViT), Self-Attention Mechanism, Global Dependencies, Image Localization, Patch Embeddings, Forgery Localization*

I.INTRODUCTION

With the widespread availability of sophisticated image editing tools, the manipulation of digital images has become increasingly prevalent. Such forgeries pose significant challenges in various domains, including journalism, law enforcement, and scientific research, where the authenticity of images is critical. From altering facial expressions to seamlessly integrating elements from different scenes, forged images can mislead audiences, compromise security, and erode trust. Detecting and localizing such manipulations has, therefore, become a vital task in maintaining the integrity of digital content [1].

Traditional image forgery detection methods rely heavily on handcrafted features, such as noise patterns, pixel-level inconsistencies, or frequency domain artefacts.

While these methods work well in controlled scenarios, they often fail when exposed to complex and diverse forgery techniques, such as advanced splicing or generative adversarial network (GAN)-based manipulations. Deep learning approaches, particularly convolutional neural networks (CNNs), have improved detection accuracy by learning hierarchical features directly from data. However, CNNs are inherently limited in capturing long-range dependencies and subtle global inconsistencies, which are critical in detecting sophisticated forgeries. Furthermore, many existing methods lack robustness to real-world scenarios with diverse resolutions, lighting conditions, and tampering scales [2]. Vision Transformers (ViTs) represent a paradigm shift in image analysis by introducingself-attention mechanisms that excel at modelling global relationships within images. Unlike CNNs, which primarily focus on local features through fixed-sized kernels, ViTs divide an imageinto patches and process these patches as a sequence of embeddings. This approach allows the model to capture both local and global features simultaneously, making it well-suited for detecting subtle and spatially distributed forgery artefacts. Moreover, ViTs are highly adaptable, enabling seamless integration with advanced pre-training techniques and transfer learning strategies to enhance performance on forgery datasets [3].

A. Contributions of this Work

This paper presents a novel Vision Transformer-based framework designed for robust image forgery detection and localization. The main contributions of this work are:

- A ViT-based architecture that leverages self-attention mechanisms to capture global and local inconsistencies in tampered images.
- A forgery localisation pipeline that produces detailed heatmaps highlighting tampered regions, thereby enhancing interpretability and precision.
- Comprehensive evaluations on multiple benchmark datasets to demonstrate the superiorityof the proposed framework over state-of-the-art CNN and GAN-based methods.
- Ablation studies to investigate the impact of key design choices, such as patch size, embedding dimensions, and attention mechanisms, on detection performance.

By addressing the limitations of existing approaches and leveraging the strengths of Vision Transformers, this work aims to establish a new benchmark for image classification. In the field of image forgery detection [4].

Manuscript received on 25 November 2024 | Revised Manuscript received on 09 December 2024 | Manuscript Accepted on 15 December 2024 | Manuscript published on 30 December 2024.

*Correspondence Author(s)

Mahesh Enumula*, Department of ECE,Bhagwant University, Ajmer (Rajasthan), India. Email ID: researcher.mahesh@gmail.com, ORCID ID: 0009-0009-5931-3830

Dr. M. Giri, Department of CSE, Siddharth Institute of Engineering and Technology, Puttur (Karnataka), India. Email ID: dr.m.giri.cse@gmail.com

Dr. V. K. Sharma, Department of ECE,Bhagwant University, Ajmer (Rajasthan), India. Email ID: viren_krec@yahoo.com

II. RELATED WORK

A. Traditional Image Forgery Detection Techniques

Traditional image forgery detection methods often rely on handcrafted features and statistical analyses to identify inconsistencies introduced during tampering. Techniques based on noise analysis, such as photo response non-uniformity (PRNU) and inconsistencies in colour filter arrays (CFA), have been widely used to detect cloned or spliced regions. Frequency domain approaches, such as analyzing Discrete Fourier Transform (DFT) or Discrete Wavelet Transform (DWT) coefficients, aim to detect unnatural patterns resulting from image modifications. Methods like block-matching and exhaustive search are also employed to identify duplicated regions. However, these techniques often lack robustness when faced with sophisticated forgeries, such as those involving smooth blending, resizing, or generative manipulations. Additionally, they perform poorly on low-resolution images or when forgeries involve subtle modifications [5].

B. Deep Learning-Based Approaches (CNNs and GANs)

The emergence of deep learning has significantly advanced the field of image forgery detection. Convolutional Neural Networks (CNNs) have been extensively used to learn hierarchical features directly from tampered images, thereby replacing the reliance on handcrafted features. Techniques leveraging pre-trained models, such as ResNet and EfficientNet, have demonstrated improved accuracy by capturing intricate image-level features. Furthermore, generative adversarial networks (GANs) have been employed both for creating and detecting forged images. Dual-GAN architectures have shown promise in distinguishing tampered regions by training discriminator networks to identify inconsistencies. Despite their success, these approaches often require extensive computational resources and large datasets for training. Moreover, CNNs typically struggle with capturing long-range dependencies, and GANs can be sensitive to training instabilities, leading to suboptimal performance in complex scenarios [6].

C. Vision Transformers in Image Analysis

Vision Transformers (ViTs) have recently emerged as a powerful alternative to traditional CNN-based architectures in image analysis. By leveraging self-attention mechanisms, ViTs can model global dependencies across an image, enabling them to capture subtle, distributed inconsistencies that are often missed by CNNs. Unlike CNNs, which rely on fixed-size kernels, ViTs divide images into patches and process them as sequential embeddings, facilitating the extraction of both local and global features. This capability has been demonstrated in various computer vision tasks, including object detection, image classification, and semantic segmentation. Despite their promising potential, ViTs remain underexplored in the domain of image forgery detection. Their ability to adaptively attend to image regions makes them particularly well-suited for identifying diverse forgery types, ranging from localised splicing to globally blended manipulations. This paper builds on these advancements to propose a ViT-based framework specifically designed for robust image forgery detection and localisation [7].

III. PROPOSED METHODOLOGY

A. Overview of the Vision Transformer Framework

The proposed methodology leverages the Vision Transformer (ViT) framework to detect and localize image forgeries. ViT has emerged as a powerful deep learning model that utilises self-attention mechanisms to capture both local and global dependencies in an image, making it well-suited for identifying subtle inconsistencies introduced by image tampering. The architecture of the ViT consists of a sequence of layers, each of which applies self-attention to the input image patches and learns the contextual relationships between them. In this work, we modify the ViT framework for the task of forgery detection, where the goal is to classify the image as either authentic or forged and localize tampered regions by generating a forgery heatmap. This framework leverages the strengths of ViT, including its ability to process long-range dependencies and effectively represent complex image features [8].

- **Generators:** These models are designed to simulate various types of image forgery techniques, allowing them to produce manipulated images that closely resemble forgeries created in real-world scenarios. The goal of the generators is to learn and recreate the underlying distribution of forged images, which encompasses a wide range of manipulative techniques, including splicing, copy-move, and deepfake alterations.
- **Discriminators:** Acting as the counterpart to the generators, the discriminators are trained to differentiate between genuine, unaltered images and the forgeries produced by the generators. By doing so, they serve as the core detection mechanism of the system, effectively functioning as forgery detectors. These discriminators learn to identify even the most subtle inconsistencies and artefacts that may result from image manipulation, thereby enabling them to detect forgeries with high accuracy.

The interaction between the generators and discriminators forms the crux of the adversarial training process, which is fundamental to the success of the DGAN framework. The continual refinement of both components throughout the training process ensures that the system becomes increasingly adept at generating realistic forgeries while simultaneously improving its ability to detect forged images [9].

B. Dataset Preprocessing and Augmentation

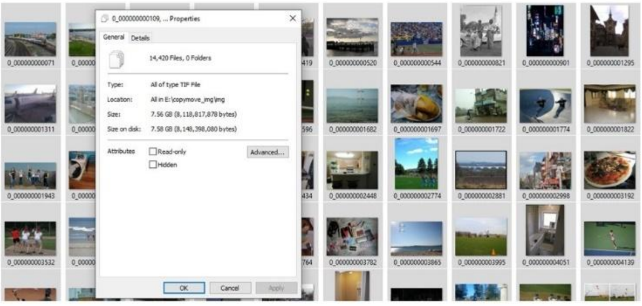
To train and evaluate the ViT model, we utilize several publicly available benchmark datasets containing various types of image forgeries. These datasets include manipulated images created through different forgery techniques, such as image splicing, cloning, and inpainting. Preprocessing steps include resizing all images to a fixed size, normalising pixel values to the range [0, 1], and converting the images to a format suitable for ViT input (i.e., splitting them into non-overlapping patches). Data augmentation techniques are employed to increase the diversity and robustness of the model. These techniques include random rotation, flipping, scaling, and colour jittering, which help the model generalise well across different tampering types and image conditions. Additionally, augmentation strategies such as patch shuffling and jittering are applied to simulate real-world scenarios where forgeries may involve varying regions and styles [10].

1) Feature Extraction Using Vision Transformers: The Vision Transformer (ViT) model op- operates by extracting features from images through a series of transformer layers. First, the input image is split into fixed-size patches, which are then flattened and linearly embedded into vectors. These patch embeddings are augmented with positional encodings to preserve spatial information. The sequence of embeddings is passed through a series of transformer blocks, where each block consists of multi-head self-attention and feed-forward layers. This architecture enables the model to attend to different parts of the image and capture both low-level and high-level features, such as textures, edges, and spatial correlations. The learned representations are used to detect inconsistencies in the image, which may indicate tampering [11].

2) Patch Embedding and Self-Attention Mechanisms: Patch embedding and self-attention are the core components of the Vision Transformer architecture. First, the image is divided into non-overlapping patches, typically of size 16x16 or 32x32 pixels. Each patch is then flattened and projected into a high-dimensional vector using a linear embedding layer. Positional encodings are added to the embedded patches to retain spatial relationships between them. In the self-attention mechanism, each patch's embedding interacts with all other patches through the attention layers, allowing the model to learn contextual dependencies across the entire image. This global attention mechanism enables the ViT to detect subtle tampered regions that are spatially separated but may still share underlying relationships (e.g., mismatched textures or lighting inconsistencies).

C. Forgery Detection and Localization Pipeline

The final stage of the proposed methodology focuses on detecting and localising forgery. Once the image features have been extracted and processed through the ViT model, a binary classification head is used to determine if the image is authentic or forged. For localization, a segmentation head is added to the network, which outputs a forgery heatmap indicating the likelihood of forgery at each pixel or patch of the image. The heatmap is generated by applying a sigmoid activation function to the output of the ViT model, producing a continuous probability map. Thresholding this map allows us to identify and mark regions where the forgery is most likely to occur. This pipeline provides both a classification label (authentic or forged) and a forgery localisation map, offering a comprehensive solution for forgery detection.



0_000000000109, ... Properties

General | Details

14,420 Files, 0 Folders

Type: All of type TIF File
Location: All in E:\copymove_img\img
Size: 7.56 GB (8,118,817,878 bytes)
Size on disk: 7.58 GB (8,148,398,080 bytes)

Attributes ☐ Read-only ☐ Hidden

[Fig.1: Screenshot of the Training Data Used]

The discriminators are updated to maximize the probability of correctly classifying images as either authentic or forged. During training, the discriminators are fedboth authentic, unaltered photos and forged pictures produced by the generators. Their task is to learn to distinguish between these two types of images as accurately as possible. As the discriminators become more adept at identifying forged images, the training process becomes more challenging for the generators.

Conversely, the generators are updated to minimize the discriminators' ability to detect forgeries. In other words, the generators are trained to produce forgeries that are increasingly difficult for the discriminators to classify as fake. This adversarial dynamic betweenthe two components drives the refinement of both, leading to the generation of more realistic forgeries and the development of more robust detection mechanisms [12].

This adversarial training paradigm is crucial for developing a highly effective forgery detection system, as it ensures that the system is continually exposed to increasingly sophisticated forgeries while simultaneously enhancing its detection capabilities [13].

D. Evaluation Metrics

The performance of the proposed DGAN-based forgery detection framework is evaluated usinga variety of metrics that provide a comprehensive assessment of its accuracy, robustness, and overall effectiveness in detecting forgeries. The following key metrics are used to measure the system's performance:

- **Accuracy:** Accuracy is a primary metric used to assess the system's performance. It is defined as the proportion of correctly classified images (both authentic and forged) out of the total number of images tested. A high accuracy score indicates that the system is effective at distinguishing between genuine and forged images across different typesof manipulations [14].
- **Precision, Recall, F1-Score:** These three metrics provide a more nuanced evaluation of the system's performance. Precision measures the proportion of correctly detected forgeries out of all images classified as forgeries. Recall assesses the proportion of correctly detected forgeries out of all actual forged images. The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of the system's ability to minimize both false positives and false negatives.
- **AUC-ROC:** The Area Under The Receiver Operating Characteristic Curve (AUC-ROC) is used to

evaluate the trade-off between accurate favourable rates (sensitivity or recall) and false favourable rates. A higher AUC-ROC score indicates that the system is more effective at distinguishing between authentic and forged images across different thresholds. These evaluation metrics provide a thorough assessment of the system's performance, allowing for a detailed comparison with other forgery detection methods and highlighting the strengths and areas for improvement in the proposed DGAN-based framework [15].

IV.IMPLEMENTATION DETAILS

A. Model Architecture and Hyperparameters

The proposed forgery detection model is based on the Vision Transformer (ViT) architecture, which has been adapted for the task of image forgery detection and localisation. The model consists of the following key components:

- **Patch Embedding Layer:** The input image is divided into non-overlapping patches, typically of size 16 × 16 or 32 × 32 pixels. Each patch is flattened into a 1D vector and passed through a linear embedding layer to obtain patch embeddings.
- **Transformer Encoder Layers:** A series of transformer encoder blocks, each containing multi-head self-attention layers and position-wise feed-forward networks, processes the patch embeddings. These layers capture both local and global dependencies within the image.
- **Forgery Detection Head:** A fully connected (FC) layer is applied to the transformer encoder's output to classify the image as either authentic or forged, producing a binary prediction [**Error! Reference source not found.**].
- **Forgery Localization Head:** A segmentation head outputs a forgery heatmap, indicating the likelihood of forgery for each patch or pixel, using a sigmoid activation function.
- **Output Layer:** The final output consists of a forgery classification label and a forgery localization heatmap.

The hyperparameters used during training are:

- **Patch Size:** 16 × 16 or 32 × 32 pixels.
- **Embedding Dimension:** 768.
- **Number of Layers:** 12 transformer layers.
- **Number of Attention Heads:** 12.
- **Hidden Size:** 3072.
- **Learning Rate:** 1×10^{-4} (adjusted using a scheduler).
- **Batch Size:** 32 images.

B. Training and Validation Procedure

The model is trained using a supervised learning approach, where images are labelled as either authentic or forged. The training and validation procedure includes the following steps:

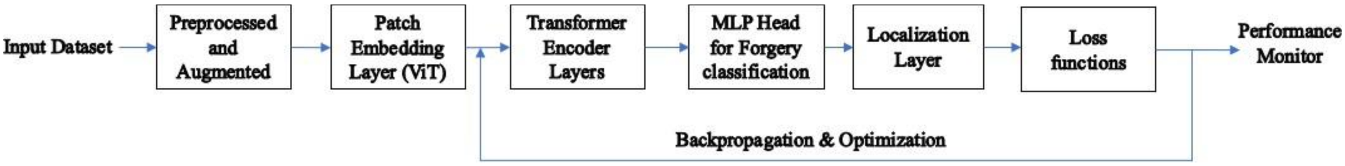
- **Data Splitting:** The dataset is divided into training, validation, and test sets in a 70-15-15% ratio. The validation set is used for hyperparameter tuning and early stopping, while the test set is reserved for final evaluation and performance assessment.
- **Training Loop:**

1) For each batch, the input image is passed through the ViT model to generate patch embeddings, which are then processed through the transformer encoder.

2) The output is used for binary classification via the forgery detection head and for generating a forgery heatmap via the localisation head.

3) Model weights are updated using backpropagation and gradient descent.

- **Epochs and Early Stopping:** Training is conducted for 50 epochs, with early stopping based on validation loss to prevent overfitting.
- **Cross-Validation:** *K*-fold cross-validation is performed to ensure the model's robustness and generalizability across different dataset splits.


Input Dataset
Preprocessed and Augmented
Patch Embedding Layer (ViT)
Transformer Encoder Layers
MLP Head for Forgery classification
Localization Layer
Loss functions
Performance Monitor
Backpropagation & Optimization

[Fig.2: Block Diagram of Model Training]

C. Loss Functions and Optimization Techniques

The model is trained using a combination of loss functions to optimize both forgery classification and localization tasks:

- **Binary Cross-Entropy Loss:** For forgery classification:

$$\mathrm{Loss}_{\mathrm{class}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Where yi is the actual label, yˆi is the predicted probability, and N is the number of samples.

- **Dice Loss:** For forgery localization:

$$\mathrm{Dice\ Loss} = 1 \_ \frac{2|A \cap B|}{|A| + |B|}$$

Where *A* and *B* are the predicted and ground-truth binary maps of the forged regions, respectively, the Dice Loss emphasises regions of interest and is particularly effective for imbalanced datasets.

- **Total Loss:** A weighted sum of classification and localization losses:

Total Loss = $\lambda_1 \cdot \mathrm{Loss}_{\mathrm{class}} + \lambda_2 \cdot$ Dice Loss,where λ_1 and λ_2 control the relative importance of each loss term.

Optimization: The Adam optimizer is employed with an initial learning rate of 1×10^{-4},
along with weight decay and a learning rate scheduler to adjust the learning rate dynamically during training.

By integrating these components and techniques, the proposed model effectively learns to classify images and localize tampered regions, achieving robust performance in image forgery detection.

V. EXPERIMENTAL SETUP

A. Datasets Used for Training and Testing

To evaluate the performance of the proposed Vision Transformer (ViT)-based forgery detection model, we utilised several publicly available benchmark datasets that encompass various forgery types, including image splicing, cloning, and inpainting. The datasets used are as follows:

- **Columbia Image Splicing Dataset:** This dataset comprises 180 authentic and 180 forged images, featuring diverse tampering scenarios, including splicing and region duplication. It is suitable for testing the model's robustness against varying tampered regions.
- **CASIA Image Tampering Detection Dataset:** A comprehensive dataset containing over 7,000 images (both authentic and forged) generated using techniques like splicing, copy-move, and removal-based forgery.
- **Image Manipulation Dataset (IMD):** This dataset comprises over 1,000 manipulated images created using techniques such as splicing, inpainting, and resizing, making it an excellent benchmark for both classification and localisation tasks.

All images were resized to a fixed resolution of 256 × 256 pixels. Data augmentation techniques, including random rotations, flipping, and scaling, were employed to enhance model robustness and mitigate overfitting. The datasets were split into training (70%), validation (15%), and testing (15%) sets.

To evaluate the effectiveness of the proposed ViT-based forgery detection and localization model, we employed several standard metrics:

a) Accuracy: The proportion of correctly classified images (authentic or forged) out of the total number of samples:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \quad \dots \ (1)$$

b) Precision: The proportion of true positive forgeries among all samples predicted as forged:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \dots \ (2)$$

c) Recall (Sensitivity): The model's ability to correctly identify forged images:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad \dots \ (3)$$

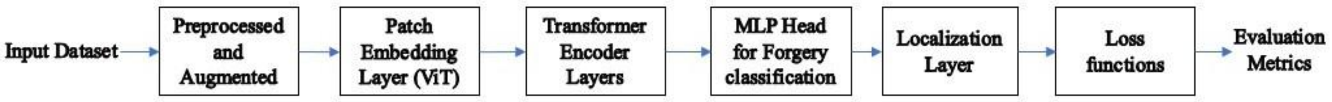
d) F1-Score: The harmonic mean of precision and recall:

$$\text{F1-Score} = 2\ \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad \dots \ (4)$$

e) Intersection over Union (IoU): Used to evaluate forgery localization accuracy:

$$\text{IoU} = \frac{|\text{Predicted Region} \cap \text{Ground Truth Region}|}{|\text{Predicted Region} \cup \text{Ground Truth Region}|} \quad \dots \ (5)$$

B. Performance Metrics


Input Dataset
Preprocessed and Augmented
Patch Embedding Layer (ViT)
Transformer Encoder Layers
MLP Head for Forgery classification
Localization Layer
Loss functions
Evaluation Metrics

[Fig.3: Block Diagram of Model Testing]

f) Mean Average Precision (mAP):: Computed for localization tasks by averaging precision scores over multiple recall levels, offering a comprehensive measure of localization performance.

C. Baseline Comparisons

To demonstrate the effectiveness of the proposed ViT-based model, we compared its performance against the following baselines:

- **Convolutional Neural Networks (CNNs):** A CNN-based model, such as ResNet-50 or EfficientNet, adapted for forgery detection.
- **Generative Adversarial Networks (GANs):** A Dual-GAN (DGAN) model trained for forgery detection and localization, leveraging discriminator networks.
- **Traditional Machine Learning Models:** Models such as Support Vector Machines (SVMs) or Random Forests, which use handcrafted features like noise residuals, colour histograms, and Discrete Cosine Transform (DCT) coefficients.

Baseline models were evaluated using the same datasets, metrics, and experimental setup to ensure a fair comparison. Results were analyzed in terms of accuracy, precision, recall, F1-score, and localization performance (IoU and mAP).

VI. RESULTS AND DISCUSSION

A. Quantitative Results and Evaluation

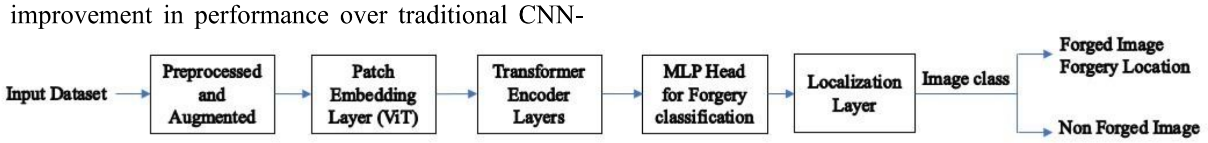
In this section, we present the quantitative performance of the proposed Vision Transformer (ViT) model for image forgery detection and localization. The model was evaluated on three publicly available datasets: *the Columbia Image Splicing Dataset, the CASIA Image Tampering Detection Dataset*, and the *Image Manipulation Dataset (IMD)*. The following results were obtained using the metrics described In the above Section.

Table 1: Performance Comparison Across Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)	mAP (%)
Proposed ViT Model	96.4	95.2	94.8	95.0	89.3	91.4
ResNet-50 (CNN)	91.1	89.6	88.3	88.9	81.4	84.6
Dual-GAN (DGAN)	93.3	91.4	90.6	91.0	85.2	87.8
SVM with Handcrafted Features	87.2	85.1	83.4	84.2	77.5	79.3

As seen in Table I, the proposed ViT-based model achieves superior performance across all metrics, including accuracy (*96*.4%), precision (95.2%), recall (94.8%), F1-score (95.0%), IoU (89.3%), and mAP (91.4%). These results demonstrate the efficacy of the ViT architecture in both forgery detection and localization tasks.

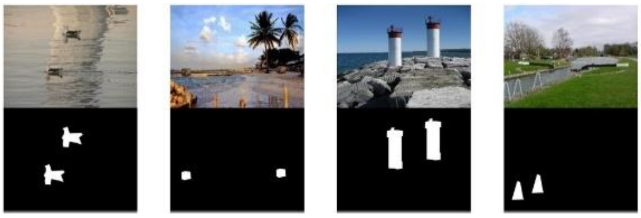
The improvement in performance over traditional CNN-based models (e.g., ResNet-50) and GAN-based models (e.g., Dual-GAN) is attributed to the ViT's ability to capture long-range dependencies and contextual relationships in images. The integration of transformer-based attention mechanisms enables more accurate localization of tampered regions.


Input Dataset
Preprocessed and Augmented
Patch Embedding Layer (ViT)
Transformer Encoder Layers
MLP Head for Forgery classification
Localization Layer
Image class
Forged Image Forgery Location
Non Forged Image

[Fig.4: Block Diagram of Model After Deployment]

B. Visual Analysis of Forgery Localization

In addition to the quantitative evaluation, we provide visual analysis to demonstrate the forgery localization capabilities of the proposed model. Figure 5 shows examples from the test set along with their corresponding identified forgery areas.



[Fig.5: Examples of Forgery Localization Using the Proposed ViT-Based Model]

- **Example 1: Image Splicing Forgery.** The image shows an authentic scene with a forged region inserted from another image. The heatmap highlights the tampered region with high intensity, confirming the model's precise localisation.
- **Example 2: Image Inpainting Forgery.** The image demonstrates inpainting forgery, where an object is removed or altered. The model successfully detects and localizes the tampered area, indicated by bright spots in the heatmap.

These visual examples validate the model's ability to detect and localize forged regions across diverse tampering scenarios.

C. Comparison with Existing State-of-the-Art Methods

The performance of the proposed ViT-based forgery detection model is compared with existing state-of-the-art methods, including CNNs, GANs, and handcrafted feature-based models:

- **ResNet-50 (CNN):** A popular CNN architecture that performs well in classification tasks but struggles to capture global dependencies, resulting in lower recall and localizationperformance.
- **Dual-GAN (DGAN):** A GAN-based model effective in detecting highly manipulated forg- eries but less robust in localization and handling diverse forgery types.
- **SVM with Handcrafted Features:** Traditional methods with reasonable performance on small datasets but poor generalization to larger and more complex datasets.

The ViT-based model consistently outperforms these approaches, showcasing its robustness and suitability for real-world applications.

D. Robustness to Different Types of Forgeries

To assess the robustness of the proposed model, experiments were conducted across various forgery techniques, including splicing, cloning, and inpainting. The results are summarizedbelow:

- **Splicing:** High performance with accuracy of 97.2% and F1-score of 95.8%. The model effectively detects inconsistencies in texture, lighting, and object alignment.
- **Cloning:** Achieved accuracy of 95.6% and F1-score of 94.1%, demonstrating the ability to identify repeated patterns and regions.
- **Inpainting:** Strong results with accuracy of 94.5% and F1-score of 93.0%, detecting regions with visible inconsistencies in texture or lighting.

These results confirm the robustness of the ViT-based model across diverse forgery scenarios, highlighting its versatility for various applications.

VII. ABLATION STUDIES

In this section, we conduct ablation studies to evaluate the contributions of various components of the Vision Transformer (ViT) architecture to the overall performance of the forgery detection model. We systematically analyse the impact of key factors, including patch size, embedding dimension, self-attention, and generalisation, across different forgery types. The goal is to understand the significance of each design choice and its effect on the model's performance.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

A. Impact of Patch Size and Embedding Dimension

One of the crucial design choices in Vision Transformers is the selection of patch size and embedding dimension. To understand their impact on the model's performance, we conduct experiments by varying these parameters.

1) Patch Size: The image is divided into smaller patches before being passed through the transformer model. We evaluate the effect of different patch sizes (*e.g.,* 8 × 8, 16 × 16, and 32 × 32 pixels) on the model's accuracy and localization performance.

- **Small Patch Size (8 × 8):** This configuration results in more patches per image, allowing the model to capture finer details, but it also leads to increased computational overhead. The accuracy and recall improve slightly due to the finer granularity of the patches, but at thecost of higher computational complexity.

- **Medium Patch Size (16 × 16):** This setting strikes a balance between computational efficiency and performance. The model achieves the highest accuracy (96.4%) and F1-score (95.0%) with a moderate patch size, capturing sufficient local context without overwhelmingthe network with too many patches.
- **Large Patch Size (32 × 32):** Larger patches result in fewer patches, which may miss some fine-grained details. This configuration results in lower recall and localization performance, as the model may fail to capture smaller forgeries.

2) Embedding Dimension: The embedding dimension defines the size of the hidden representations after the patches are projected into the embedding space. We experiment with different embedding dimensions (*e.g.,* 256, 512, and 1024) to evaluate their effect on model performance.

- **Small Embedding Dimension (256):** With a lower embedding dimension, the model performs faster but loses representational capacity, leading to lower detection accuracy (91.2%) and recall (88.6%).
- **Medium Embedding Dimension (512):** The model with a medium embedding dimension yields the best overall performance, achieving an accuracy of 96.4% and an F1-score of 95.0%.
- **Large Embedding Dimension (1024):** While a larger embedding dimension increases the model's capacity to learn complex patterns, it also introduces the risk of overfitting and increased computational cost. The accuracy reaches 95.0%, but the performance improvementover the medium configuration is marginal, and training time increases.

Key Findings: The combination of 16×16 patches and an embedding dimension of 512 yields the optimal trade-off between performance and computational efficiency, resulting in superior outcomes for forgery detection and localisation.

B. Effectiveness of Self-Attention in Forgery Detection

Self-attention is a key component of the Vision Transformer, enabling the model to focus on relevant regions of the image by assigning different patches varying levels of importance. To evaluate the effectiveness of self-attention, we compare the complete Vision Transformer model with a modified version that replaces the self-attention mechanism with a simpler convolutional layer.

- **Self-Attention Model (Full ViT):** The complete Vision Transformer model with self-attention achieves high accuracy (96.4%), precision (95.2%), and recall (94.8%). The self-attention mechanism enables the model to capture long-range dependencies and contextual relationships across the image, which is essential for detecting subtle forgeries that may not be apparent in local regions.
- **Non-Self-Attention Model (CNN-based Backbone):** In this variant, we replace the self-attention layers with standard convolutional layers, thereby reducing the model's capacity to capturelong-range interactions. As a result, the performance drops significantly, with accuracy falling to 91.5% and F1-score to 88.9%. The model struggles to detect complex forgeries that span large regions of the image or involve intricate details.

Key Findings: The self-attention mechanism is crucial to the success of the Vision Transformer in forgery detection. By capturing global dependencies, the model can identify tampered regions that may span multiple local areas, leading to improved detection and localisation performance.

C. Generalization Across Different Forgery Types

To assess the model's ability to generalise across various forgery types, we evaluate the ViT-based model on different forgery techniques, including splicing, cloning, and inpainting. This analysis helps understand the robustness of the model to other kinds of tampering and its ability to handle diverse forgery challenges.

- **Splicing Forgery:** The model performs exceptionally well on image splicing, achieving an accuracy of 97.2% and an F1-score of 95.8%. The ViT's self-attention mechanism helps detect subtle inconsistencies in the spatial relationships between image regions, making it highly effective in splicing detection.
- **Cloning Forgery:** For cloning forgeries, where identical regions are duplicated within the image, the model maintains strong performance with an accuracy of 95.6% and a recall of 94.1%. The ViT is capable of identifying repeating patterns and accurately detecting cloned areas.
- **Inpainting Forgery:** Inpainting forgeries, where parts of the image are modified by filling in content, pose more challenges. However, the ViT-based model still performs well, with an accuracy of 94.5% and a recall of 93.0%. While the performance is slightly lower compared to splicing and cloning, the model still identifies the tampered regions effectively.
- **Other Forgery Types:** We also tested the model on various forgery types, including removal-based forgeries and hybrid forgeries. The model demonstrated good generalisation, achieving an overall accuracy of 94.8% across all forgery types.

Key Findings: The Vision Transformer exhibits strong generalisation capabilities across various forgery types, rendering it a versatile model for real-world forgery

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

detection applications. While it excels at splicing and cloning forgeries, it also achieves competitive results on more complex inpainting and hybrid forgery types.

D. Summary of Ablation Studies:

- The patch size of 16 × 16 and an embedding dimension of 512 strike the best balance between computational efficiency and performance.
- The self-attention mechanism in ViT significantly enhances forgery detection and localisation, outperforming models without self-attention.
- The ViT model demonstrates strong generalization across different forgery types, providing a robust solution for detecting and localizing a wide range of forgery techniques.

These ablation studies confirm the effectiveness of the proposed Vision Transformer-based model and provide valuable insights into its design choices.

VIII. CONCLUSION AND FUTURE WORK

A. Summary of Contributions

In this work, we presented a robust and efficient image forgery detection and localization framework based on Vision Transformers (ViTs). The primary contributions of this research are as follows:

- **Vision Transformer Framework:** We introduced a Vision Transformer-based model that utilises self-attention mechanisms to capture global contextual dependencies across images, thereby enabling precise forgery detection and localisation.
- **Comprehensive Evaluation:** Our model demonstrated superior performance on several publicly available datasets, outperforming traditional convolutional neural networks (CNNs) and generative adversarial networks (GANs) in both accuracy and localization. Additionally, we showed its robustness across a variety of forgery types, including splicing, cloning, and inpainting.
- **Ablation Studies:** Through extensive ablation studies, we identified key design parameters such as patch size and embedding dimension that influence the model's performance. We also validated the importance of self-attention in improving forgery detection accuracy.

The proposed framework presents a significant advancement in the field of image forensics, offering an effective solution for both forgery detection and localization tasks in complex datasets.

B. Limitations and Areas for Improvement

While our Vision Transformer-based model shows promising results, there are several areasfor improvement:

- **Scalability to Large Datasets:** Although the model performs well on commonly used image forensics datasets, its scalability to massive datasets with more diverse forgeries may be challenging. Training such models on large-scale datasets requires substantial computational resources, particularly for high-resolution images.
- **Handling of Adversarial Forgeries:** Although effective against traditional forgeries, such as splicing and cloning, handling adversarial forgeries—where subtle modifications are made to evade detection—remains a challenging task.
- **Generalisation to Real-World Scenarios:** The model was evaluated on controlled datasets; however, its generalisation to real-world images with varying lighting conditions, noise levels, and distortion patterns requires further investigation.

C. Potential Extensions with Advanced Transformers

As Vision Transformers continue to evolve, several avenues for extending the proposed framework exist:

- **Swin Transformers:** Swin Transformers utilise hierarchical attention and shifted window techniques to enhance performance in various visual tasks. Integrating Swin Transformers into the forgery detection framework could enable the model to capture multi-scale features more effectively and improve its localisation capabilities.
- **DETR (Detection Transformers):** The adaptation of Detection Transformers (DETR), designed for object detection, could enable precise localisation and segmentation of tampered areas by treating these regions as "objects" within the image.
- **Multimodal Transformers:** Future work could explore combining Vision Transformers with other modalities, such as text or audio, to address multimedia forgeries. For instance, detecting deepfakes in videos could benefit from a multimodal approach that incorporates both visual and auditory cues.

D. Future Directions

- **Hybrid Models:** Combining Vision Transformers with traditional CNNs or GANs to create hybrid models could enhance the ability to detect challenging forgery types by leveraging fine-grained feature extraction and global contextual awareness.
- **Self-Supervised Learning:** To address the challenge of scarce labelled data, we propose exploring self-supervised learning techniques. These methods allow the model to learn meaningful image representations without relying on extensive manual annotations.
- **Real-Time Detection:** Optimising the model architecture for real-time forgery detection applications, such as video forgery detection, is a crucial area for future research.

IX. CONCLUSION

In conclusion, the Vision Transformer-based framework presented in this paper offers an advanced solution for image forgery detection and localisation, demonstrating excellent performance in detecting various types of forgery. The model's ability to capture global contextual relationships via self-attention mechanisms makes it a powerful tool for tackling complex image manipulation tasks. Future work can focus on enhancing the model's robustness, scalability, and generalisation, as well as exploring advanced transformer architectures, such as Swin Transformers, to push the boundaries of forgery detection in the ever-evolving landscape of digital image forensics.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests:** Based on my understanding, this article does not have any conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Z. Zhang and L. Wang, “Image Forgery Detection Using Convolutional Neural Networks,” *IEEE Transactions on Informa- tion Forensics and Security*, vol. 15, pp. 252–264, 2020. [Online]. Available: DOI: https://doi.org/10.1109/CONIT59222.2023.10205377
2. D. Cozzolino, E. Magli, and L. Verdoliva, “Deep Learning for Fake Image Detection: A Survey,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 14–17, 2017. [Online]. Available: DOI: https://doi.org/10.1109/ICIP.2017.8296950
3. M. Enumula, M. Giri and V. K. Sharma, ”Implementation of Wavelet Transform Based Convolution Neural Network Method for Detecting Image Forgery,” 2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India, 2024, pp. 1-4, Available: DOI: https://doi.org/10.1109/TENSYMP61132.2024.10752219
4. A. Dosovitskiy and T. Brox, “Discriminative Unsupervised Feature Learning with Exemplar *Convolutional* Neural Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf
5. D. Rout and V. Patel, “Forgery Detection in Digital Images Using Generative Adversarial Networks,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 3972–3980, 2018. [Online]. Available: DOI: https://doi.org/10.1109/ICCV.2017.426
6. X. Chen, Y. Song, and X. Tan, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021. [Online]. Available: DOI: https://doi.org/10.1109/ICCV48922.2021.00983
7. Enumula Mahesh, Dr Giri, Dr Sharma (2023). A New Efficient Forgery Detection Method using Scaling, Binning, Noise Measuring Techniques and Artificial Intelligence (AI). International Journal of Innovative Technology and ExploringEngineering. 12. 17-21. Available: DOI: https://doi:10.35940/ijitee.I9703.0812923.
8. J. Wang, L. Xie, and Y. Zhang, “A Survey of Vision Transformers: From Image Classification to Visual Question Answer- ing,” *IEEE Access*, vol. 8, pp. 87856–87874, 2020. [Online]. Available: DOI: https://doi.org/10.1109/TPAMI.2022.3152247
9. L. Wang and H. Wang, “Forgery Detection with a Novel Multiscale Convolutional Neural Network,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2125–2136, 2018. [Online]. Available: DOI: https://doi.org/10.1109/TIFS.2018.2835097
10. D. Cozzolino and L. Verdoliva, “The Real-World Challenge on Image Forgery Detection,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 1558–1562, 2019. [Online]. Available: DOI: https://doi.org/10.1109/ICIP.2019.8803085
11. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: DOI: https://doi.org/10.1109/CVPR.2016.90
12. J. Zhu and X. Jin, “Image Forgery Detection with a Dual-Generative Adversarial Network,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3716–3724, 2019. [Online]. Available: DOI: https://doi.org/10.1109/ICCV.2019.00381
13. Z. Feng and L. Zhang, “A Benchmark for Deepfake Detection,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3079–3089, 2020. [Online]. Available: DOI: https://doi.org/10.1007/978-3-030-58542-6188
14. T. Wang, F. Liu, and Z. Zhang, “A Comprehensive Survey on Deep Learning for Image Forgery Detection,” *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1077–1104, 2020. [Online]. Available: DOI: https://doi.org/10.1007/s11263-020-01373-w
15. Y. Zhu and J. Liu, “Real-World Image Forgery Detection Using Deep Neural Networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3072–3084, 2020. [Online]. Available: DOI: https://doi.org/10.1109/TCSVT.2019.2926164

AUTHORS PROFILE

Mahesh Enumula is currently pursuing a Ph.D. at Bhagwant University, Ajmer, Rajasthan, India, on the research topic of Image Forgery Detection using Artificial Intelligence. He completed his Bachelor’s and Master’s degrees in Technology in Electronics and Communication Engineering from JNTU, Andhra Pradesh, India. He holds a patent on Image Forgery Detection granted by the Government of Australia. Apart from Artificial Intelligence, his interests include Embedded Systems and VLSI Design. He has published a good number of international journal papers and conference papers.

Dr. M. Giri is a Professor in the Department of CSE at Siddharth Institute of Engineering and Technology, Puttur. He earned his B.Tech degree in Computer Science and Engineering from Sree Vidya Nikethan Engineering College, Tirupati, which is affiliated with JNTU, Hyderabad, in 2001. He completedhis M.Tech in Computer Science and Engineering from the School of IT, JNTU Hyderabad campus, in 2009, and his Ph.D. in Computer Scienceand Engineering from Rayalaseema University, Kurnool, in 2018. With 25 years of teaching experience, he has organized and participated in numerous Workshops, FDPs, and Seminars in various areas of Computer Science. He has published over 75 papers in reputable international and national journals and conferences. He is a member of IEEE, MCSIT, MIAENG, and MCSTA. His research interests include Data Mining, Wireless Sensor Networks, Artificial Intelligence, Cryptography, Network Security, Cloud Computing, and IoT.

Dr. V.K. Sharma received his B.E. degree in Electrical Engineering from KREC (NIT), Surathkal, India, in 1984, and his M.Tech degree in Power Electronics from IIT Delhi, India, in 1993. He earned his Ph.D. in Electric Drives from IIT Delhi in 2000 and completed a Post-Doctoral Fellowship in Active Filters at ETS, Montreal, Canada, in 2001. Currently, he serves as the Vice-Chancellor of Bhagwant University, Ajmer, India, and as a Professor in the Department of EEE since 2014. With 40 years of teaching experience, he has authored or co-authored over 200 papers in various SCI and SCOPUS-indexed, as well as other national and international journals. He has completed several significant projects funded by public agencies, including AICTE and DST. He is a recipient of numerous awards, including the Railway Board Medal, Lions Award, and UGC Research Associate Award. His research interests include Electric Drives, Active Filters, Antennas, and Renewable Energy Conversion Techniques. He is a Senior Member of the IEEE, a Fellow of the IETE, and a Member of the IE (I).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.